Uniwersytet Ekonomiczny w Katowicach

Text mining
Grupowanie

Bogna Zacny

Wydział Informatyki i Komunikacji
Katedra Inżynierii Wiedzy

zima 19/20

- Wprowadzenie
- Grupowanie hierarchiczne
- Grupowanie niehierarchiczne

- Wprowadzenie
- Grupowanie hierarchiczne
- Grupowanie niehierarchiczne

Grupowanie

Grupowanie obiektów polega na znajdowaniu skończonego zbioru klas (podzbiorów) w bazie danych.
Celem grupowania jest podział zbioru na stosunkowo homogeniczne (jednorodne, zgodne) grupy (klasy) zwane **klastrami** (skupieniami) różniące się względem siebie.

Grupowanie - etapy

- wybór cech,
- wybór miary odległości,
- grupowanie i interpretacja.

Grupowanie - miara odległości

Miara odległości musi spełniać następujące warunki (aksjomaty):

- $d(x, y) = 0 \Leftrightarrow x = y$,

Grupowanie - miara odległości

Miara odległości musi spełniać następujące warunki (aksjomaty):

- $d(x, y) = 0 \Leftrightarrow x = y$,
- $d(x, y) = d(y, x)$ (warunek symetrii),

Grupowanie - miara odległości

Miara odległości musi spełniać następujące warunki (aksjomaty):

- $d(x, y) = 0 \Leftrightarrow x = y$,
- $d(x, y) = d(y, x)$ (warunek symetrii),
- $d(x, z) \leqslant d(x, y) + d(y, z)$ (warunek trójkąta).

Grupowanie - miary odległości

Wśród miar odległości wyróżniamy:

- miara euklidesowa $d_{euc}(d_l, d_k) = \sqrt{\sum_{i=l}^{n}(d_{li} - d_{ki})^2}$,

Grupowanie - miary odległości

Wśród miar odległości wyróżniamy:

- miara euklidesowa $d_{euc}(d_l, d_k) = \sqrt{\sum_{i=l}^{n}(d_{li} - d_{ki})^2}$,
- miara Manhattan (miejska) $d_{+}(d_l, d_k) = \sum_{i=1}^{n} |d_{li} - d_{ki}|$,

Grupowanie - miary odległości

Wśród miar odległości wyróżniamy:

- miara euklidesowa $d_{euc}(d_l, d_k) = \sqrt{\sum_{i=l}^{n}(d_{li} - d_{ki})^2}$,
- miara Manhattan (miejska) $d_{+}(d_l, d_k) = \sum_{i=1}^{n} |d_{li} - d_{ki}|$,
- miara Czebyszewa (nieskończoności) $d_{\infty}(d_l, d_k) = \max_{i=1}^{n} |d_{li} - d_{ki}|$.

Grupowanie - miary odległości

Do miar odległości zalicza się miary oparte na miarach podobieństwa:

- miara cosinusowa

$$sim_{cos}(d_l, d_k) = \frac{d_l \cdot d_k}{|d_l|\,|d_k|} = \frac{\sum_{i=1}^{n}(d_{li}d_{ki})}{\sqrt{\sum_{i=1}^{n}(d_{li})^2}\sqrt{\sum_{i=1}^{n}(d_{ki})^2}}$$

$$d_{cos}(d_l, d_k) = 1 - sim_{cos}(d_l, d_k),$$

Grupowanie - miary odległości

Do miar odległości zalicza się miary oparte na miarach podobieństwa:

- miara cosinusowa

$$sim_{cos}(d_l, d_k) = \frac{d_l \cdot d_k}{|d_l|\,|d_k|} = \frac{\sum_{i=1}^{n}(d_{li}d_{ki})}{\sqrt{\sum_{i=1}^{n}(d_{li})^2}\sqrt{\sum_{i=1}^{n}(d_{ki})^2}}$$

$$d_{cos}(d_l, d_k) = 1 - sim_{cos}(d_l, d_k),$$

- miara Jaccarda

$$sim_{Jacc}(d_l, d_k) = \frac{|d_l \cap d_k|}{|d_l \cup d_k|}$$

$$d_{Jacc}(d_l, d_k) = 1 - sim_{Jacc}(d_l, d_k).$$

Grupowanie - metody

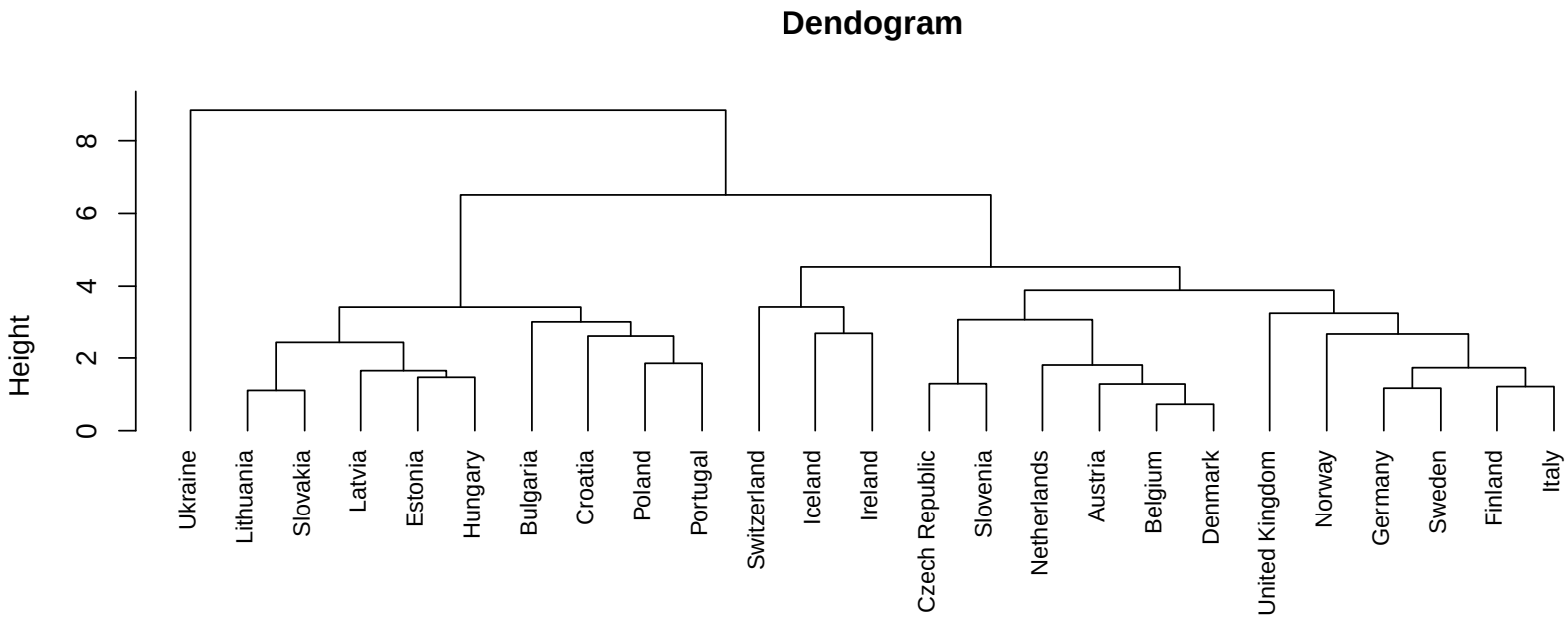
Podstawowy podział metod grupowania jako kryterium przyjmuje mechanizm grupowania, wyróżniamy:

- metody hierarchiczne,
- metody niehierarchiczne (iteracyjno-optymalizacyjne).

- Wprowadzenie
- Grupowanie hierarchiczne
- Grupowanie niehierarchiczne

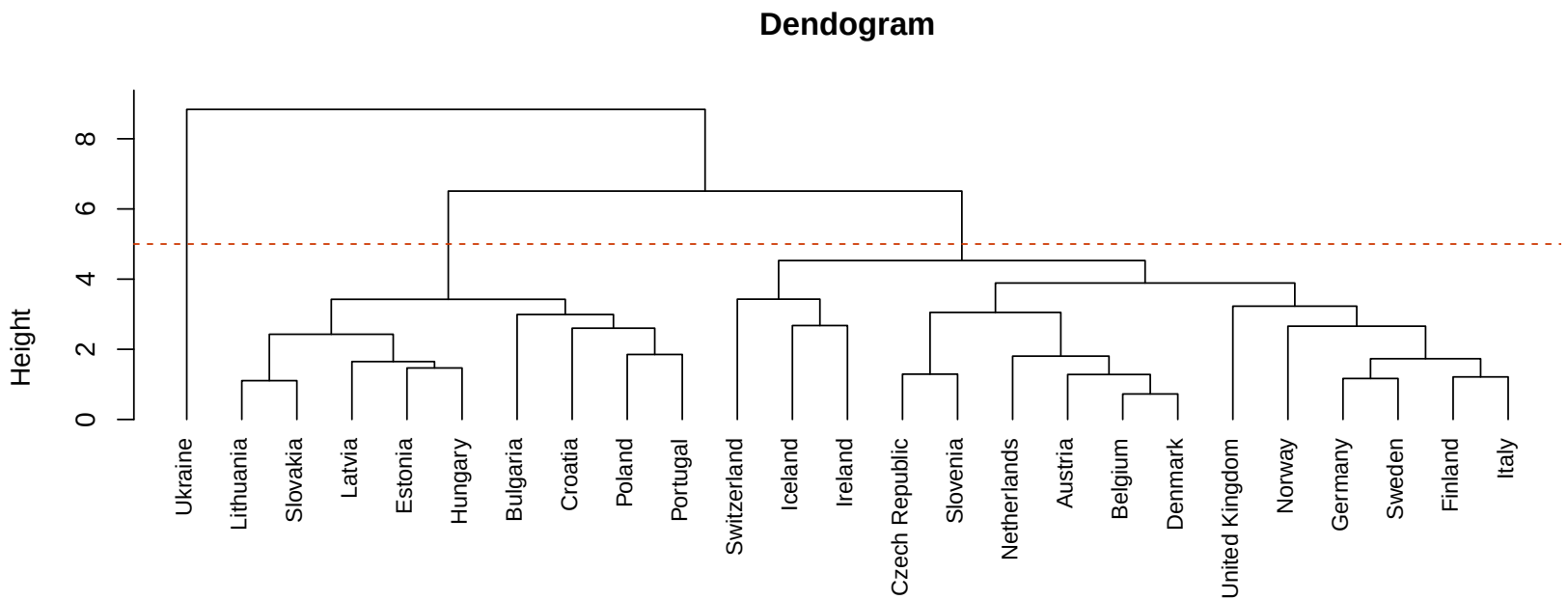
Metody hierarchiczne

Generują sekwencję podzbiorów zbioru, w wyniku której otrzymywany jest dendrogram (struktura drzewiasta).

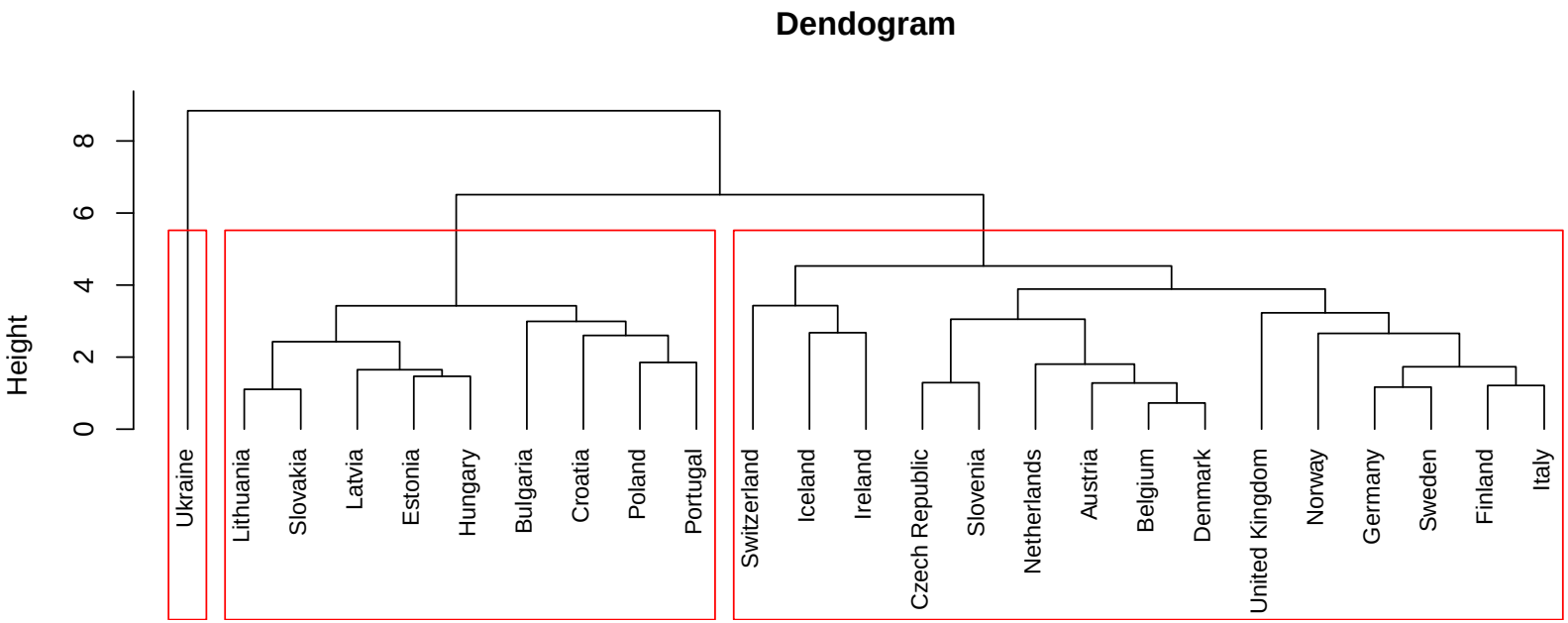


Dendogram

Grupy państw Unii Europejskiej podzielone ze względu na cechy: Area, GDP, Inflation, Life.expect, Military, Pop.growth, Unemployment



Dendogram


Dendogram
Height
0
2
4
6
8
Ukraine
Lithuania
Slovakia
Latvia
Estonia
Hungary
Bulgaria
Croatia
Poland
Portugal
Switzerland
Iceland
Ireland
Czech Republic
Slovenia
Netherlands
Austria
Belgium
Denmark
United Kingdom
Norway
Germany
Sweden
Finland
Italy

Metody hierarchiczne

Wyróżniamy dwie podmetody:

- metody aglomeracyjne – w pierwszym kroku zakłada się, że każda obserwacja stanowi jedną grupę, w kolejnych krokach dwie grupy, które są najbliżej łączone są w nową wspólną grupę, ostatecznie wszystkie rekordy należą do jednej (obejmującej wszystkie elementy) grupy;

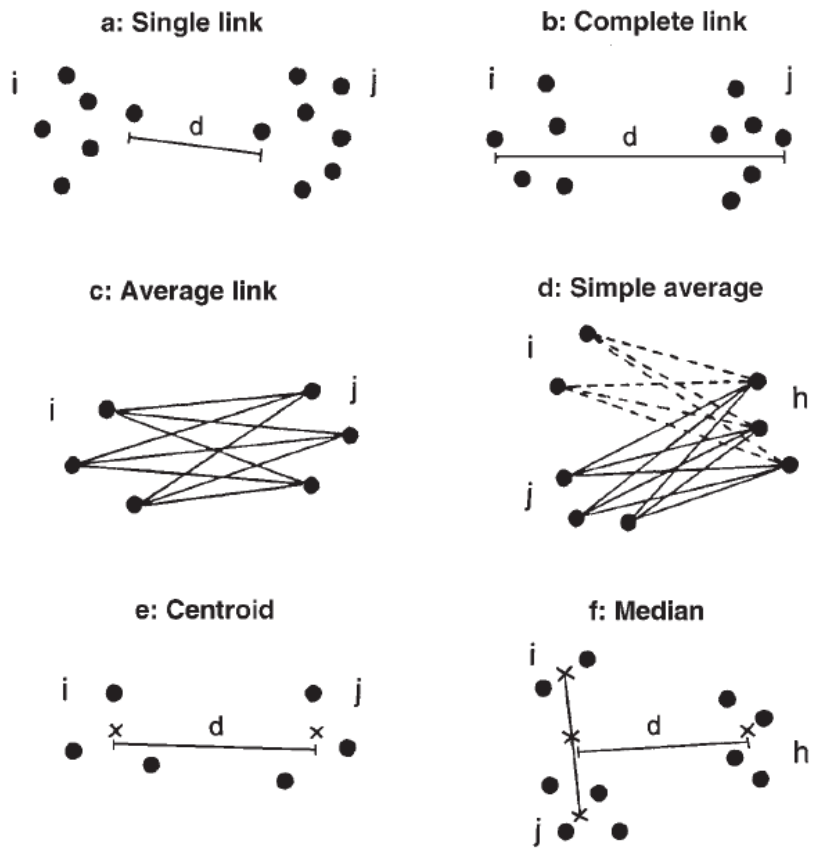
Metody hierarchiczne

Wyróżniamy dwie podmetody:

- metody aglomeracyjne – w pierwszym kroku zakłada się, że każda obserwacja stanowi jedną grupę, w kolejnych krokach dwie grupy, które są najbliżej łączone są w nową wspólną grupę, ostatecznie wszystkie rekordy należą do jednej (obejmującej wszystkie elementy) grupy;
- metody rozdzielające – w pierwszym kroku zakłada się, że wszystkie obserwacje stanowią jedną grupę, w kolejnych krokach najbardziej niepodobne rekordy wyodrębniane są i rozdzielane w osobne grupy, ostatecznie każdy rekord reprezentuje osobną grupę.

Metody hierarchiczne - odległości między skupieniami

Kluczowym zagadnieniem, obok określenia odległości pomiędzy poszczególnymi rekordami, jest określenie odległości pomiędzy grupami. Wyróżniamy klika kryteriów określania tej odległości:

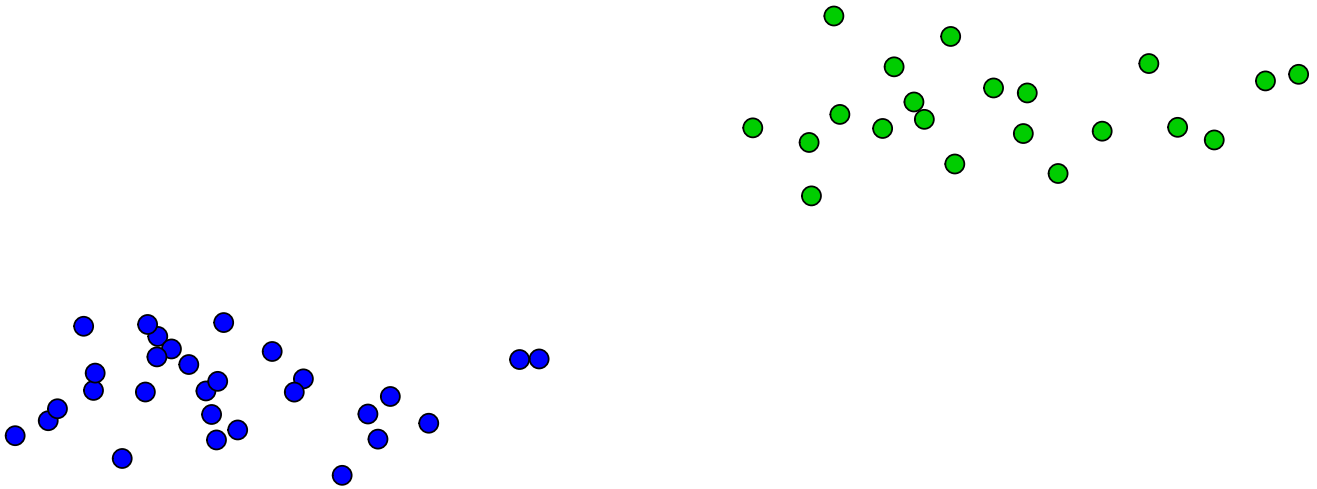


Metody hierarchiczne - odległości między skupieniami

Najczęściej wykorzystywanymi są:

- metoda pojedynczego połączenia,
- metoda całkowitego połączenia,
- metoda średniego połączenia.

Metody hierarchiczne - odległości między skupieniami

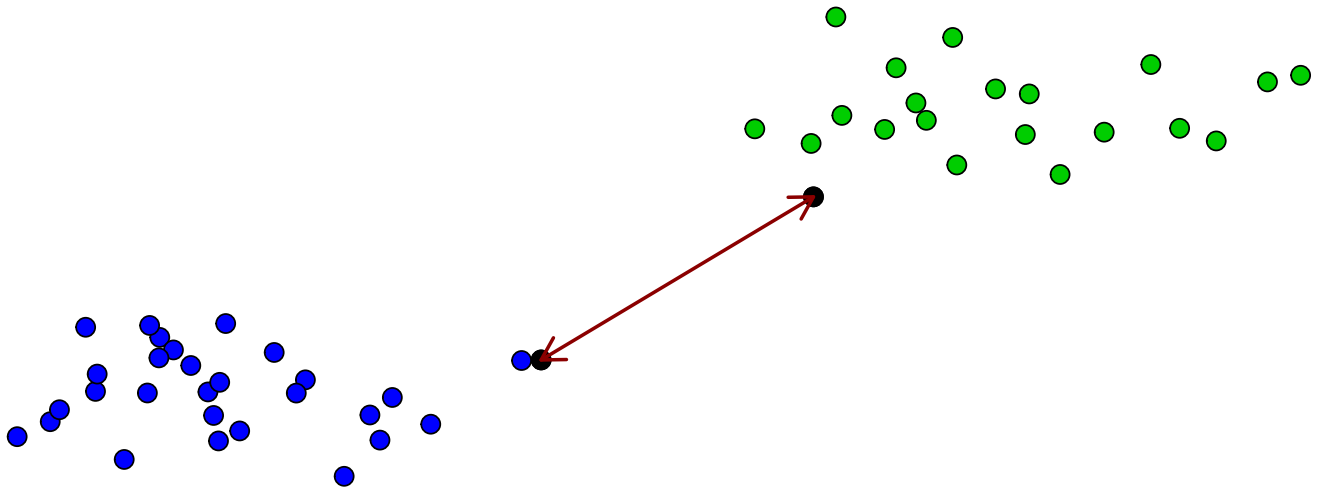


Metody hierarchiczne - odległości między skupieniami

Metoda pojedynczego połączenia – najbliższego sąsiedztwa
odległość pomiędzy skupieniami jest zdefiniowana jako odległość miedzy dwoma najbliższymi punktami, po jednym z każdej grupy. Poszukiwana jest minimalna odległość pomiędzy dowolnymi rekordami z dwóch grup.

Metody hierarchiczne - odległości między skupieniami

Metoda pojedynczego połączenia – najbliższego sąsiedztwa

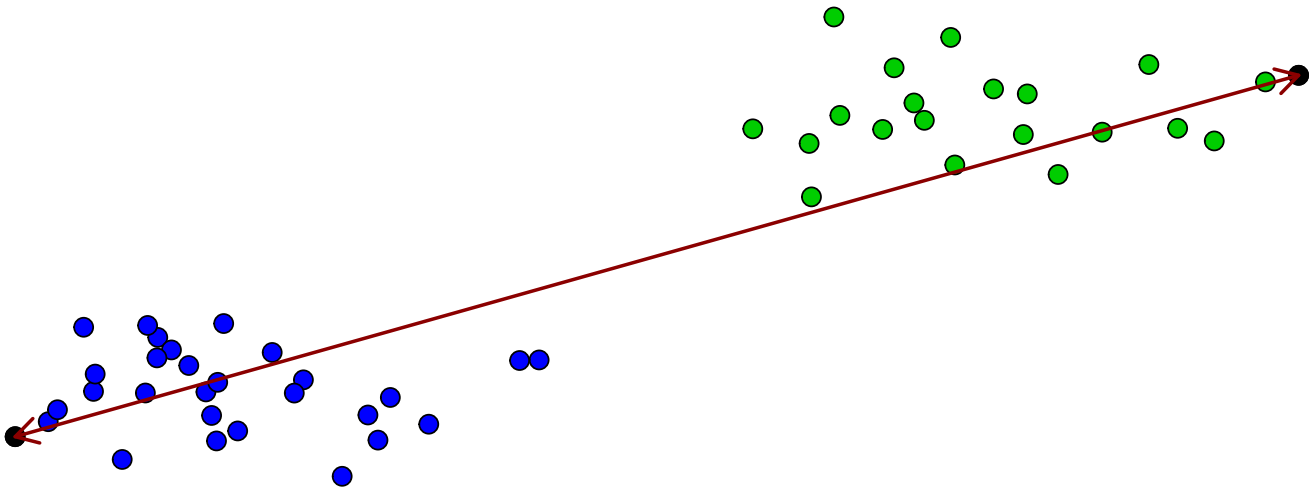


Metody hierarchiczne - odległości między skupieniami

Metoda całkowitego połączenia – najdalszego sąsiedztwa
odległość pomiędzy skupieniami jest zdefiniowana jako odległość miedzy dwoma najbardziej oddalonymi punktami, po jednym z każdej grupy. Poszukiwana jest minimalna odległość pomiędzy dowolnymi rekordami z dwóch grup, które są najbardziej oddalone od siebie.

Metody hierarchiczne - odległości między skupieniami

Metoda całkowitego połączenia – najdalszego sąsiedztwa

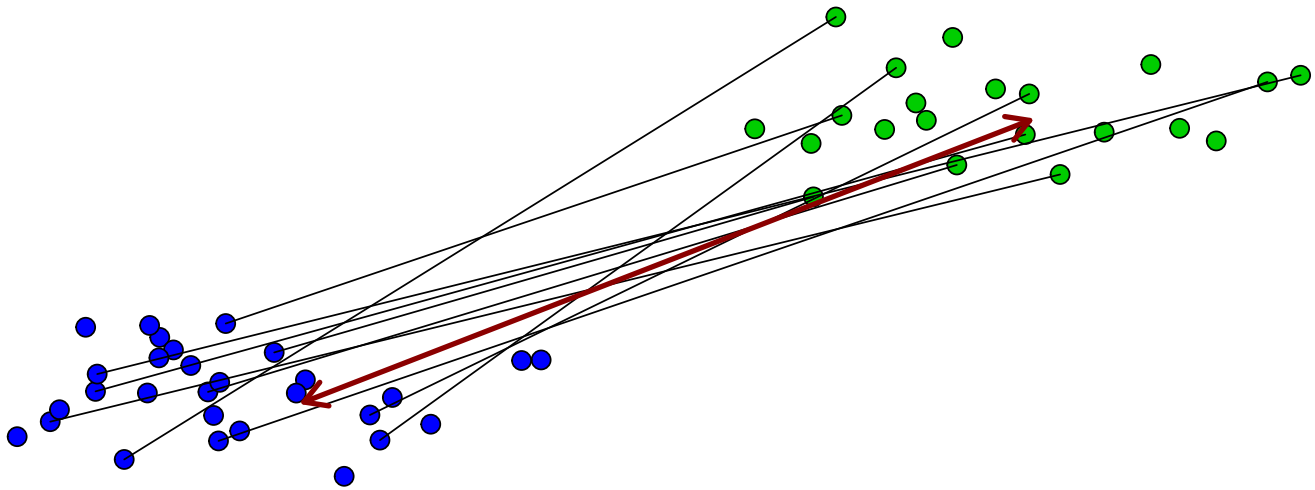


Metody hierarchiczne - odległości między skupieniami

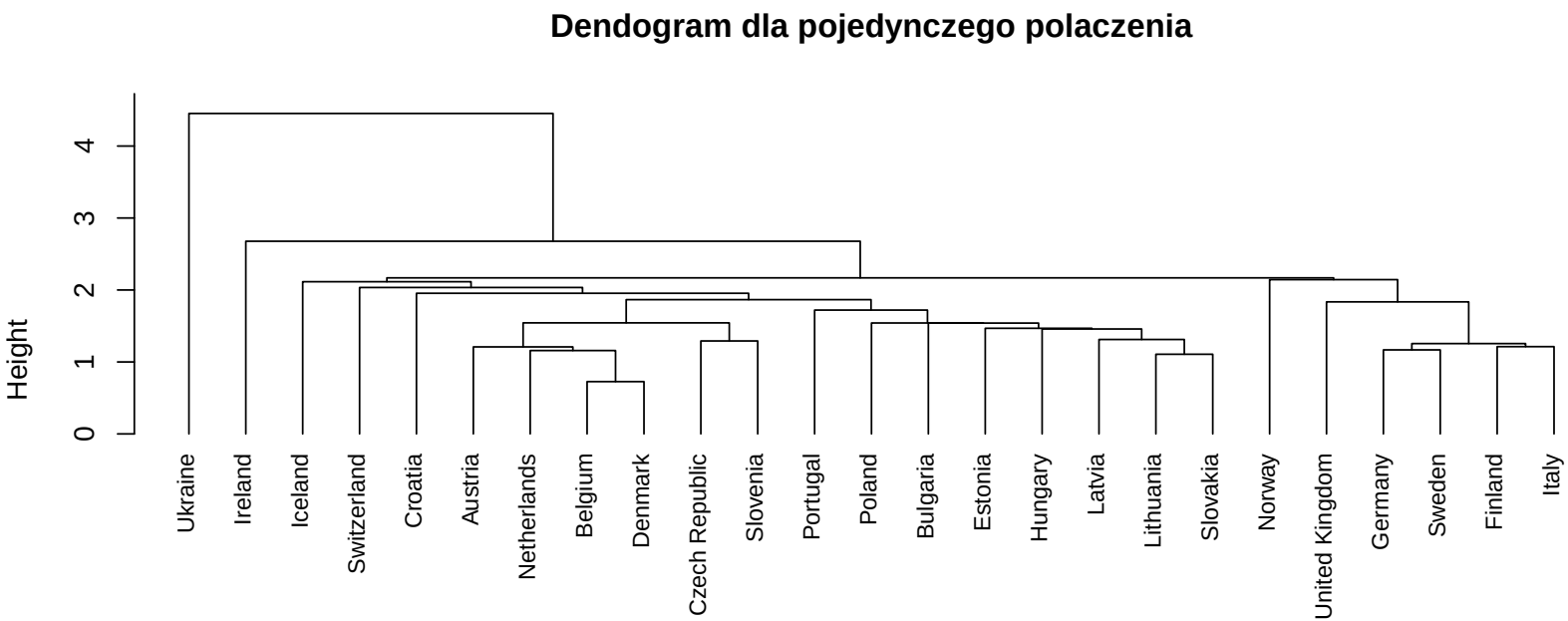
Metoda średniego połączenia
odległość pomiędzy skupieniami jest zdefiniowana jako średnia odległość wszystkich rekordów z poszczególnych grup. Poszukiwana jest minimalna wartość średniej.

Metody hierarchiczne - odległości między skupieniami

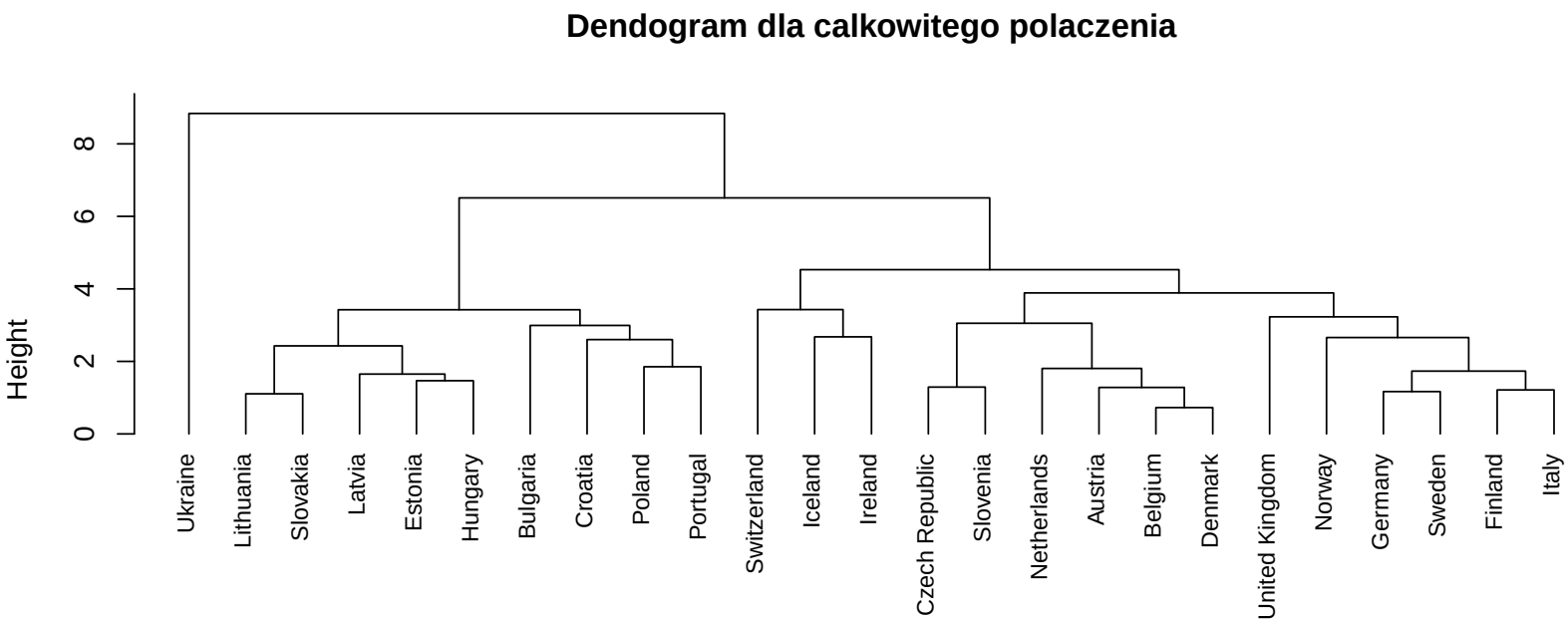
Metoda średniego połączenia



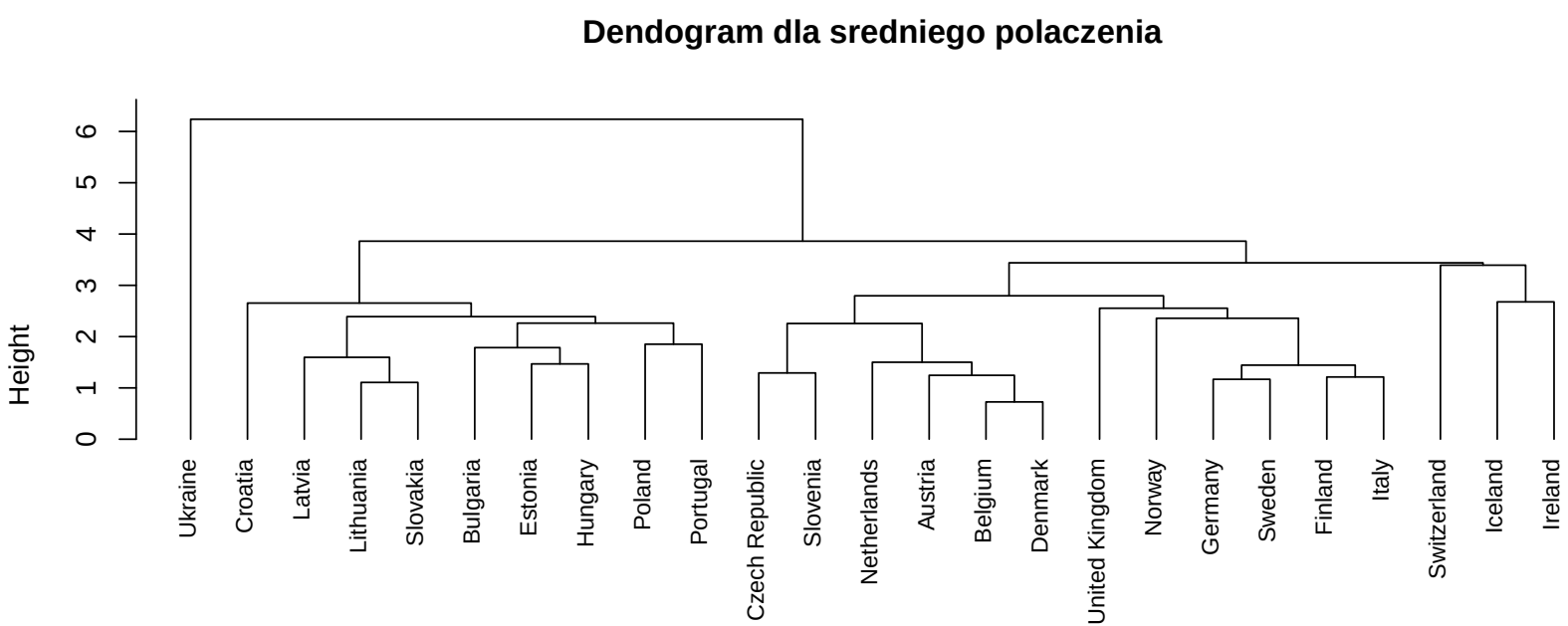
Metody hierarchiczne - odległości między skupieniami



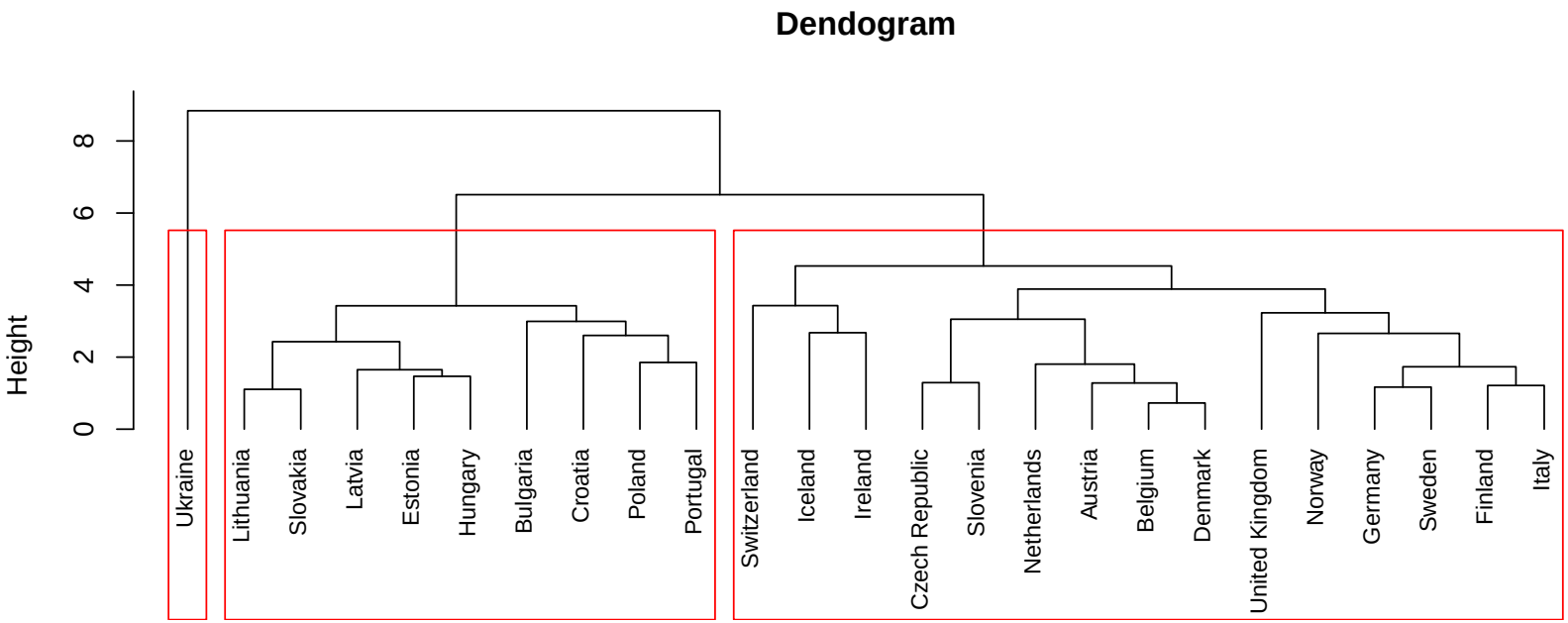
Metody hierarchiczne - odległości między skupieniami



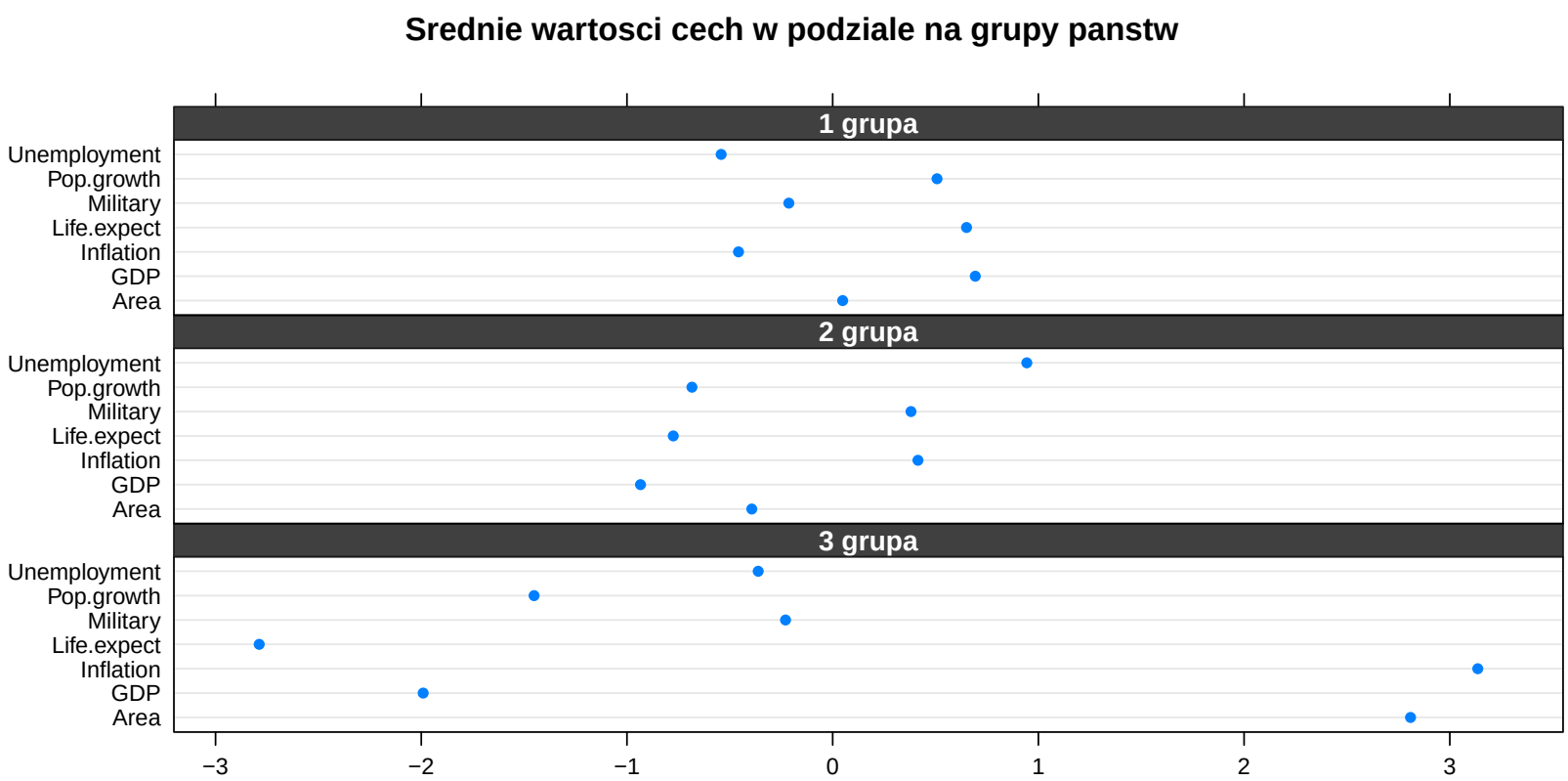
Metody hierarchiczne - odległości między skupieniami



Metody hierarchiczne - interpretacja



Metody hierarchiczne - interpretacja



- Wprowadzenie
- Grupowanie hierarchiczne
- Grupowanie niehierarchiczne

Metody iteracyjno-optymalizacyjne

Cechą charakterystyczną tej grupy metod jest zadana z góry (przez analityka) liczba skupień.
Tworzony jest początkowy podział obiektów a następnie, stosując technikę iteracyjnej realokacji obiektów pomiędzy klastrami, podział ten jest modyfikowany w taki sposób, aby uzyskać poprawę podziału zbioru obiektów pomiędzy klastry.

Algorytm *k-średnich*

Algorytm realizowany jest w 3 krokach:

- w kroku pierwszym, wybieranych jest losowo k obiektów jako początkowe środki k klastrów;

Algorytm *k-średnich*

Algorytm realizowany jest w 3 krokach:

- w kroku pierwszym, wybieranych jest losowo k obiektów jako początkowe środki k klastrów;
- w kroku drugim, obiekty alokowane są do klastrów. Każdy obiekt jest przydzielany do tego klastra, dla którego odległość obiektu od środka klastra (centroidu) jest najmniejsza;

Algorytm *k-średnich*

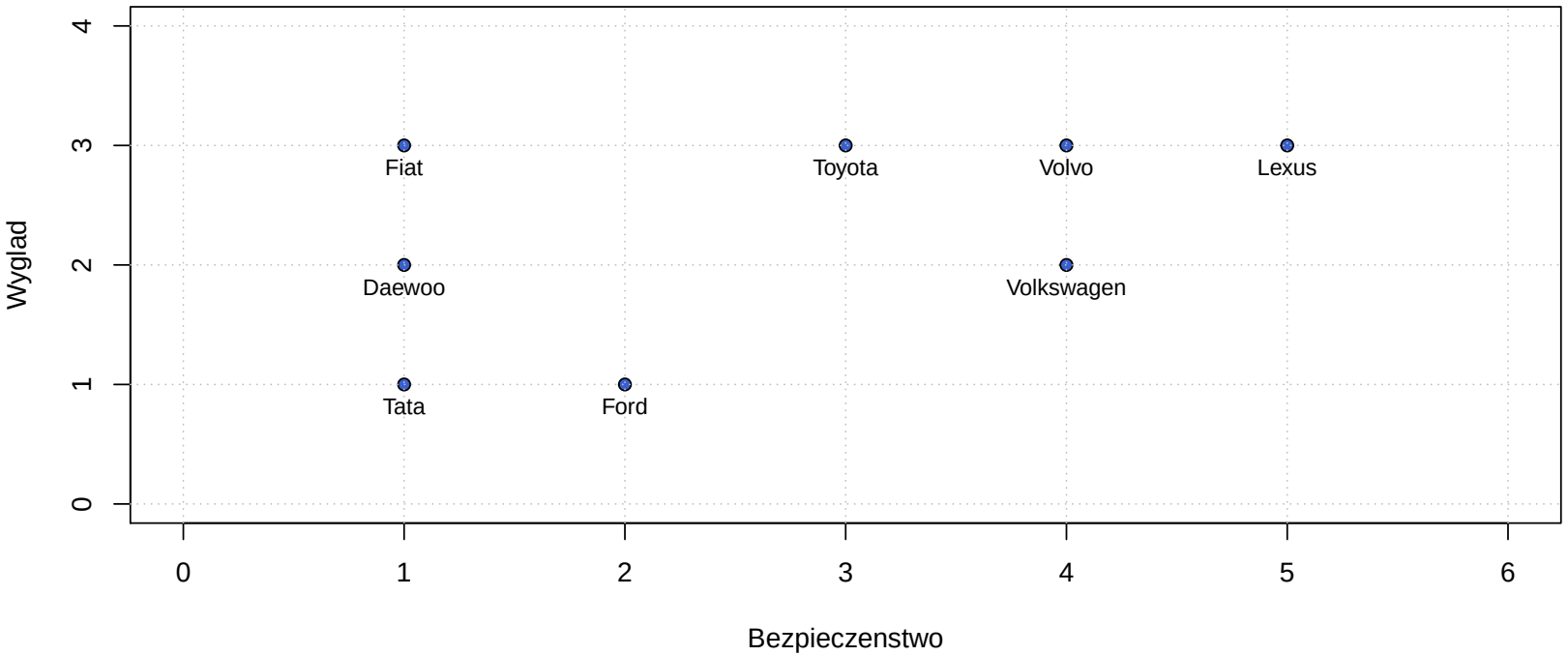
Algorytm realizowany jest w 3 krokach:

- w kroku pierwszym, wybieranych jest losowo k obiektów jako początkowe środki k klastrów;
- w kroku drugim, obiekty alokowane są do klastrów. Każdy obiekt jest przydzielany do tego klastra, dla którego odległość obiektu od środka klastra (centroidu) jest najmniejsza;
- w kroku trzecim, po alokacji obiektów do klastrów, uaktualniane są wartości średnie klastrów (środki klastrów) i ponownie wracamy do kroku alokacji obiektów do klastrów;

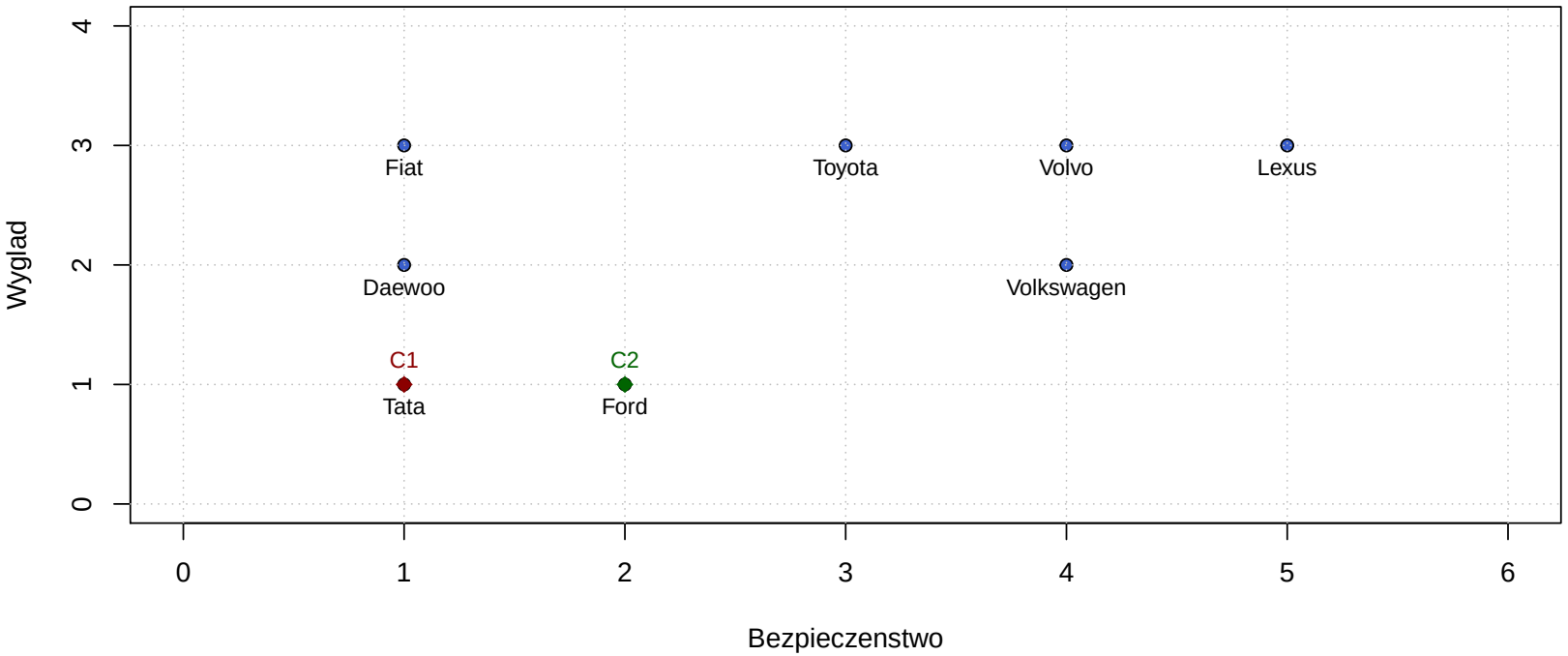
Algorytm *k-średnich* - przykład

```
##            Bezpieczeństwo Wygląd
## Daewoo                  1      2
## Fiat                    1      3
## Ford                    2      1
## Lexus                   5      3
## Tata                    1      1
## Toyota                  3      3
## Volkswagen              4      2
## Volvo                   4      3
```

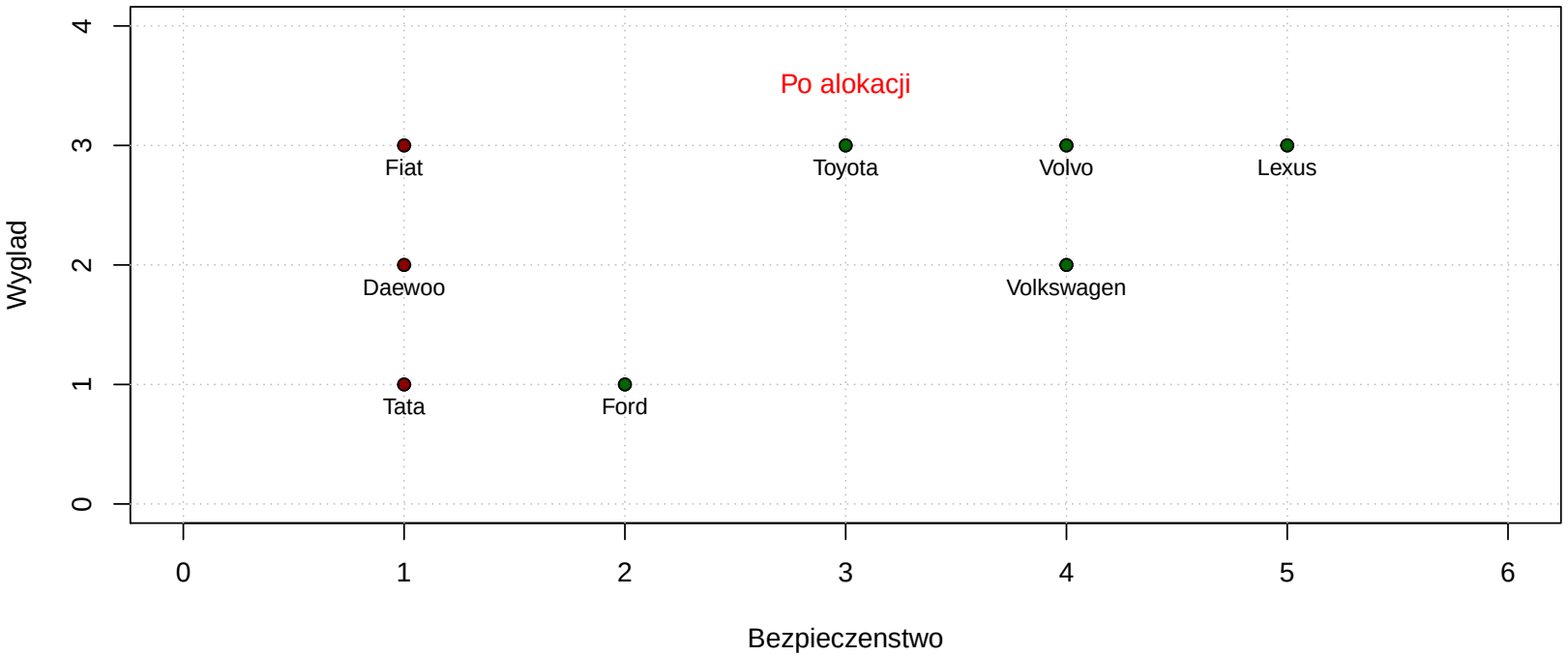
Algorytm *k-średnich* - przykład



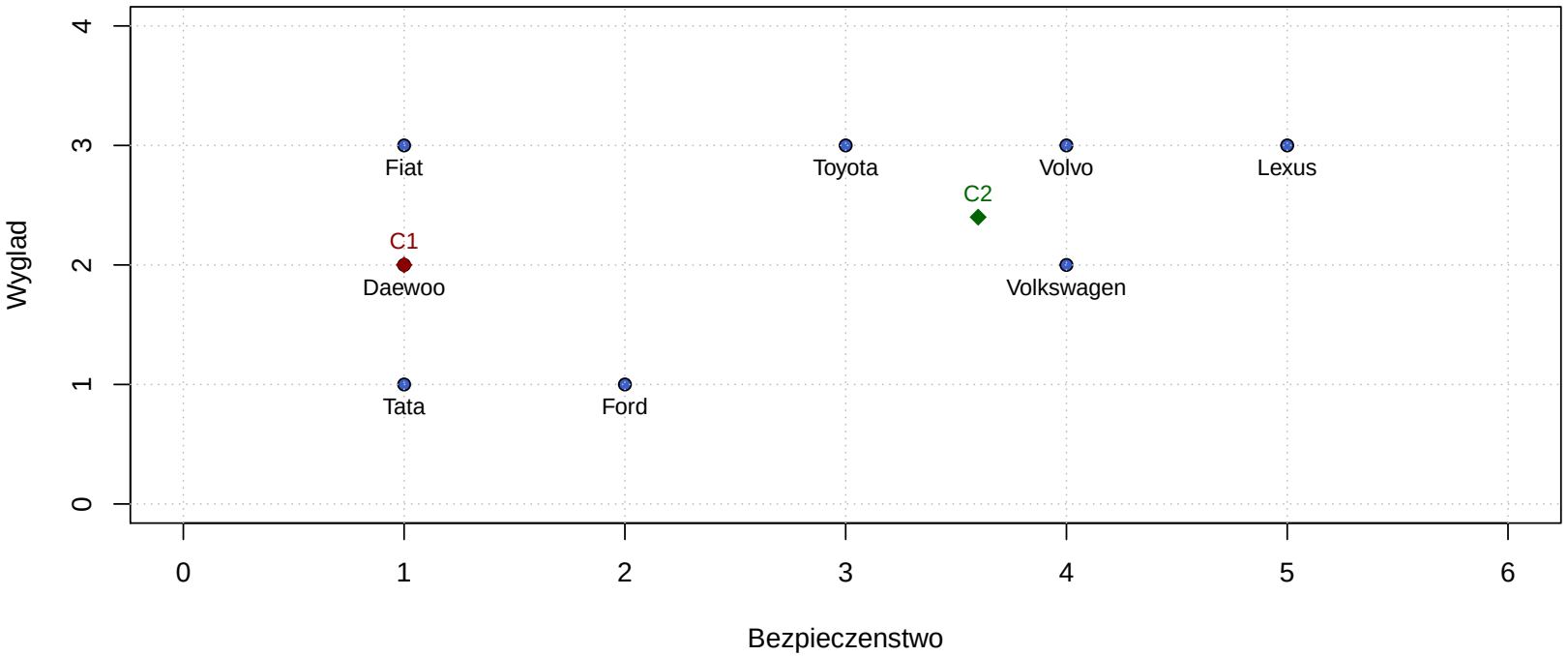
Algorytm *k-średnich* - 1. iteracja 1. krok



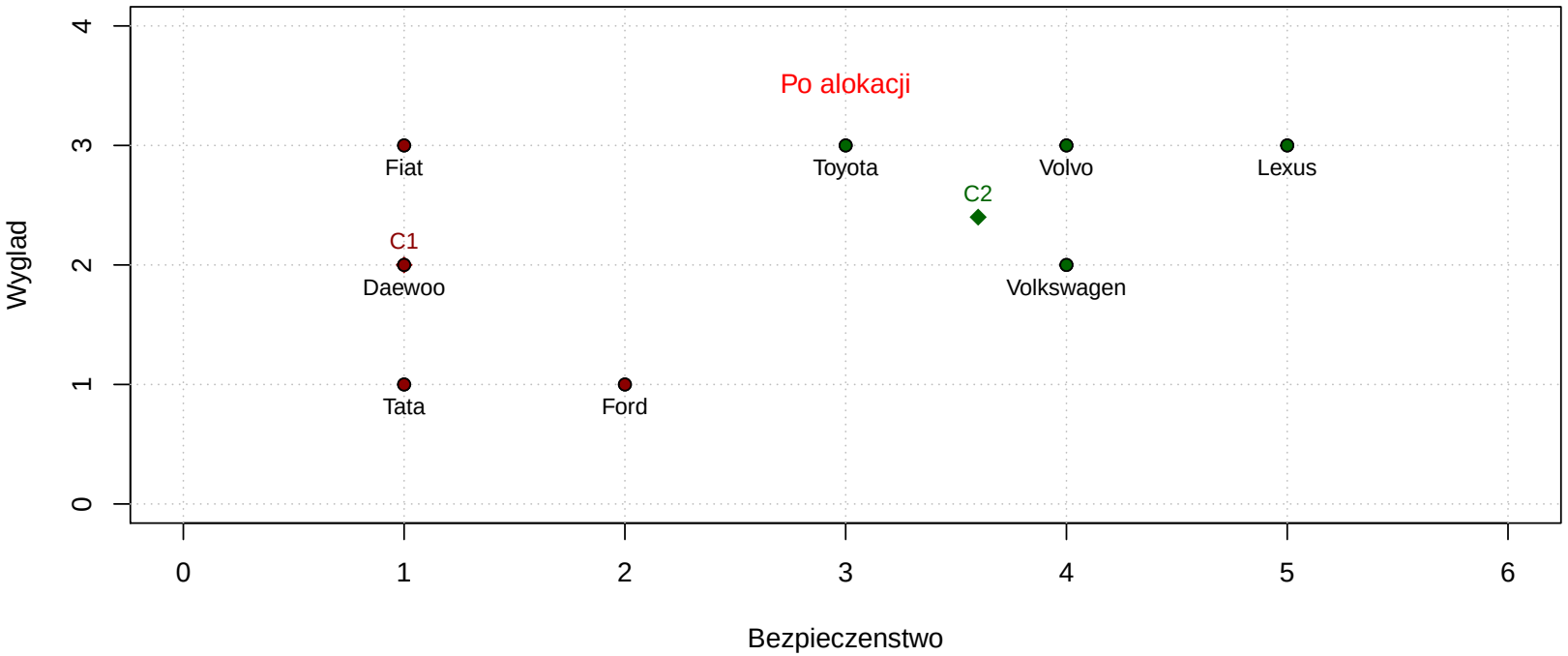
Algorytm *k-średnich* - 1. iteracja 2. krok



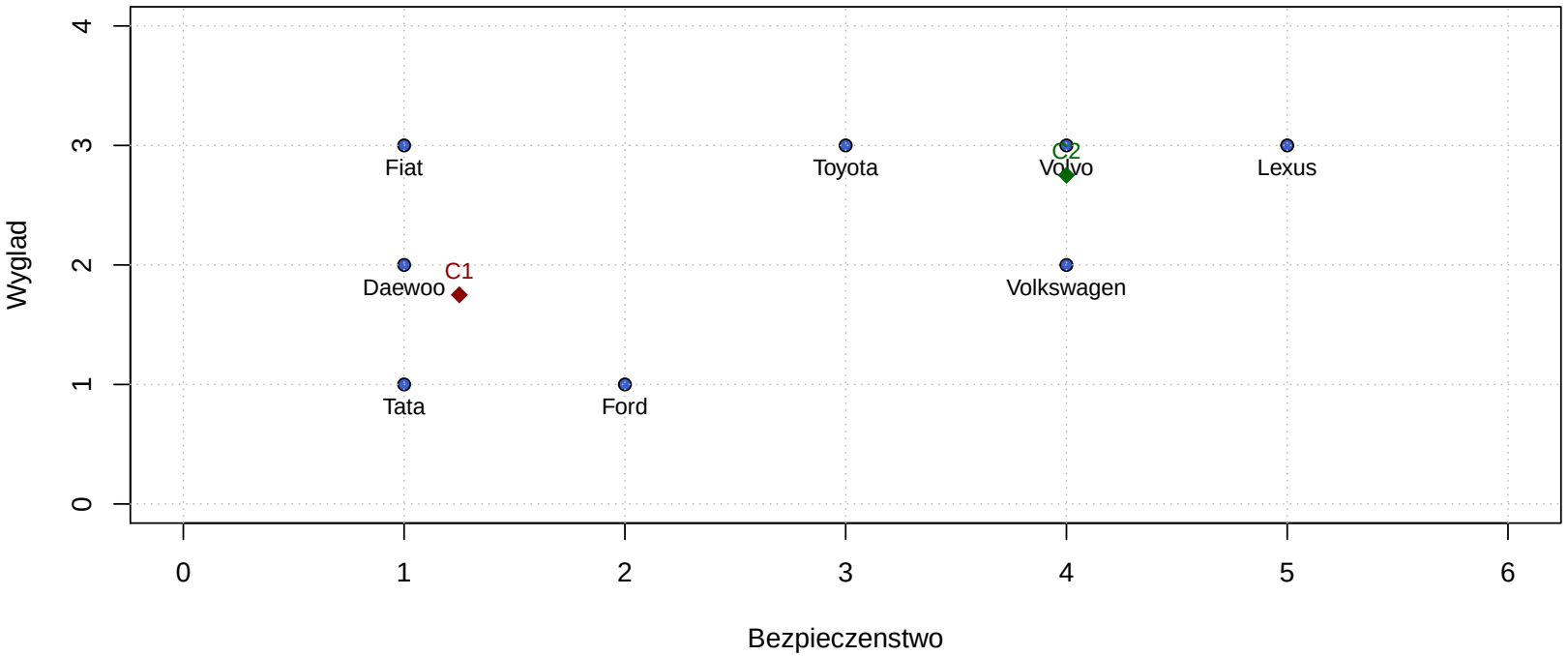
Algorytm *k-średnich* - 1. iteracja 3. krok



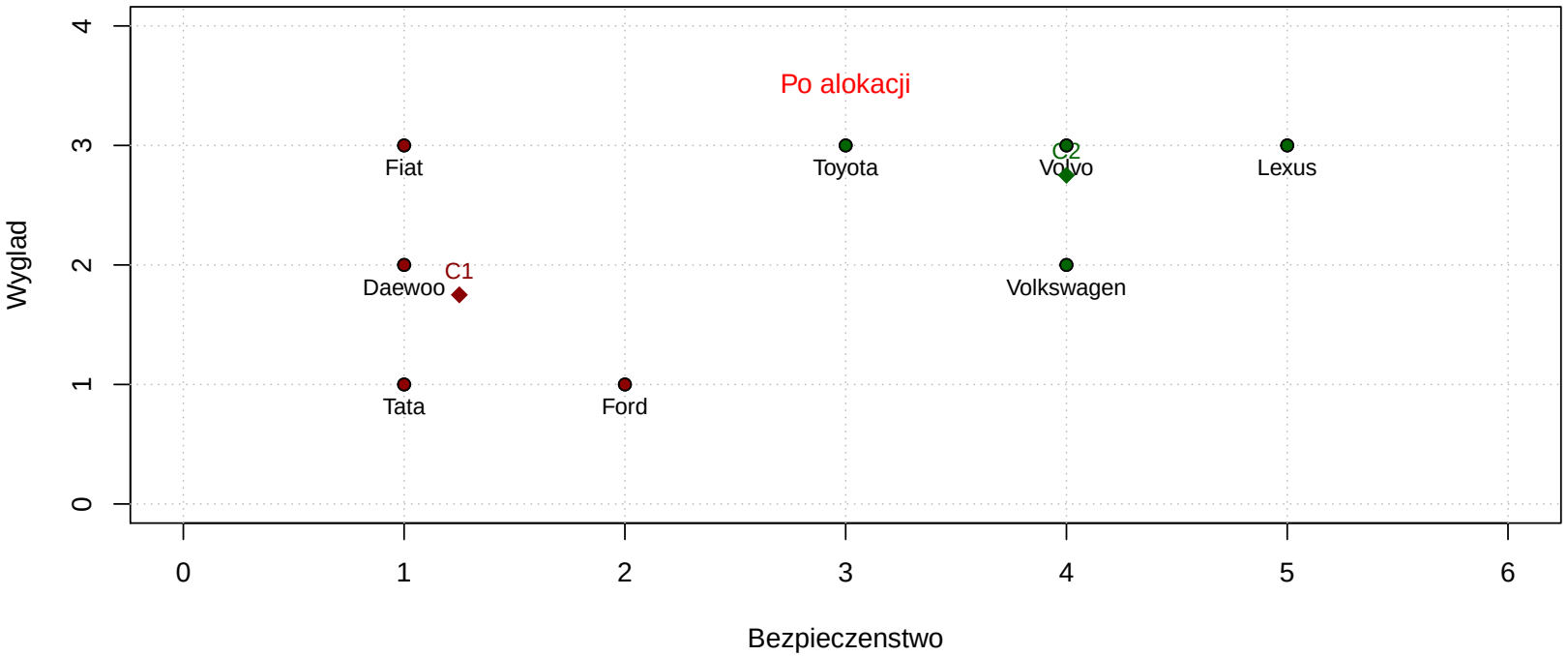
Algorytm *k-średnich* - 2. iteracja 2. krok



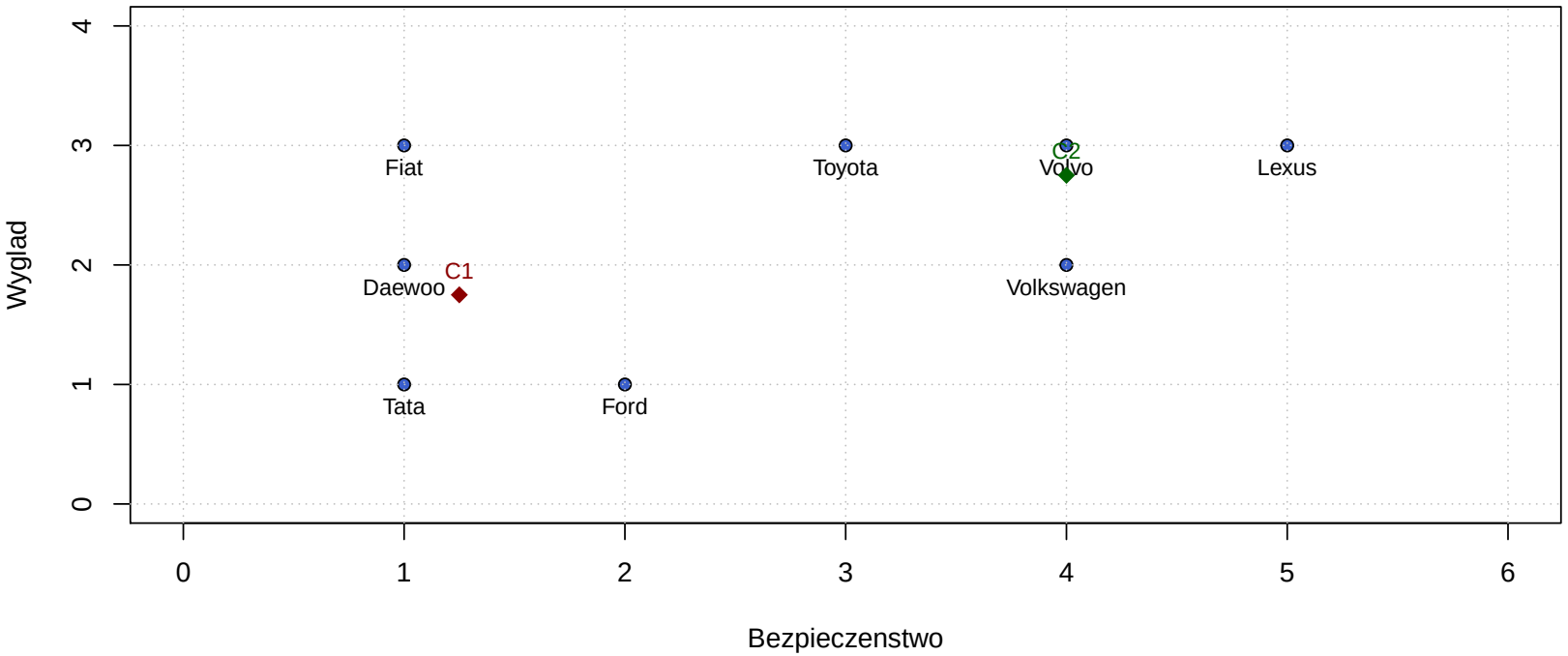
Algorytm *k-średnich* - 2. iteracja 3. krok



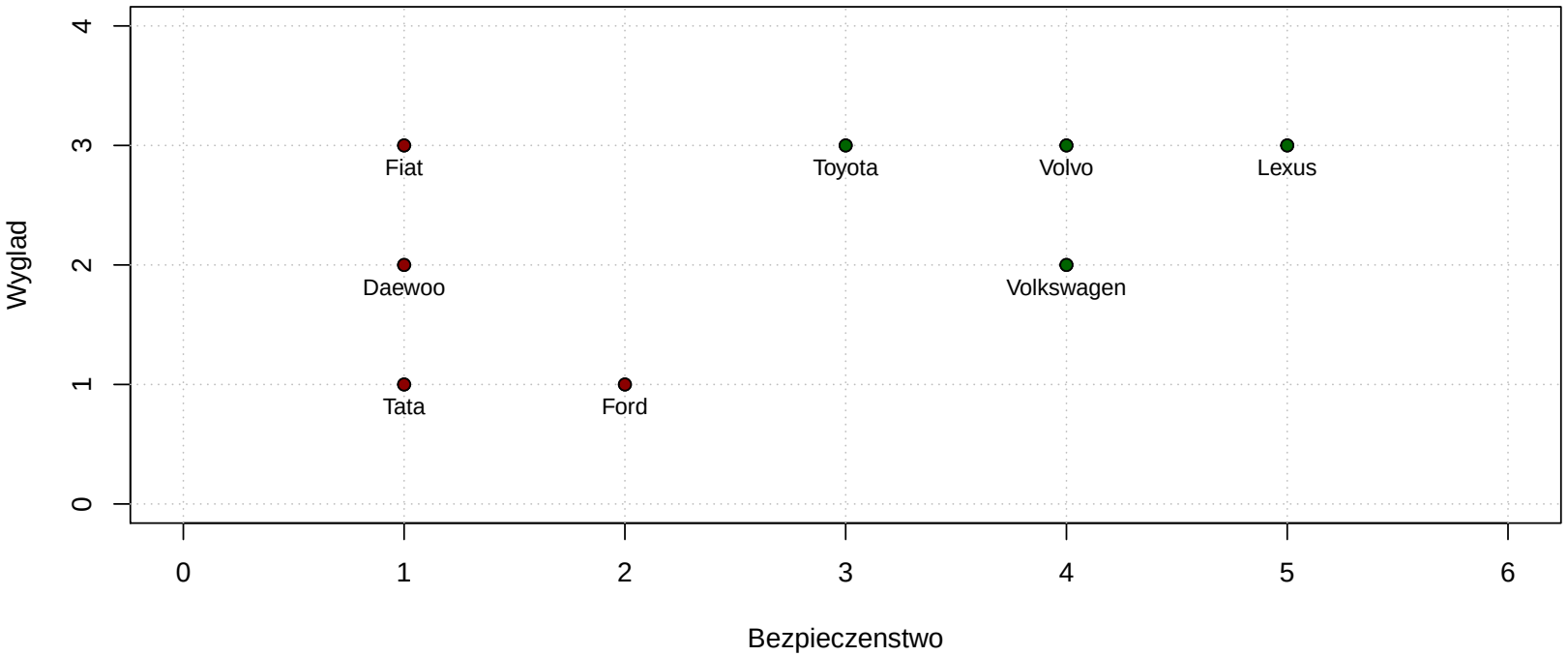
Algorytm *k-średnich* - 3. iteracja 2. krok



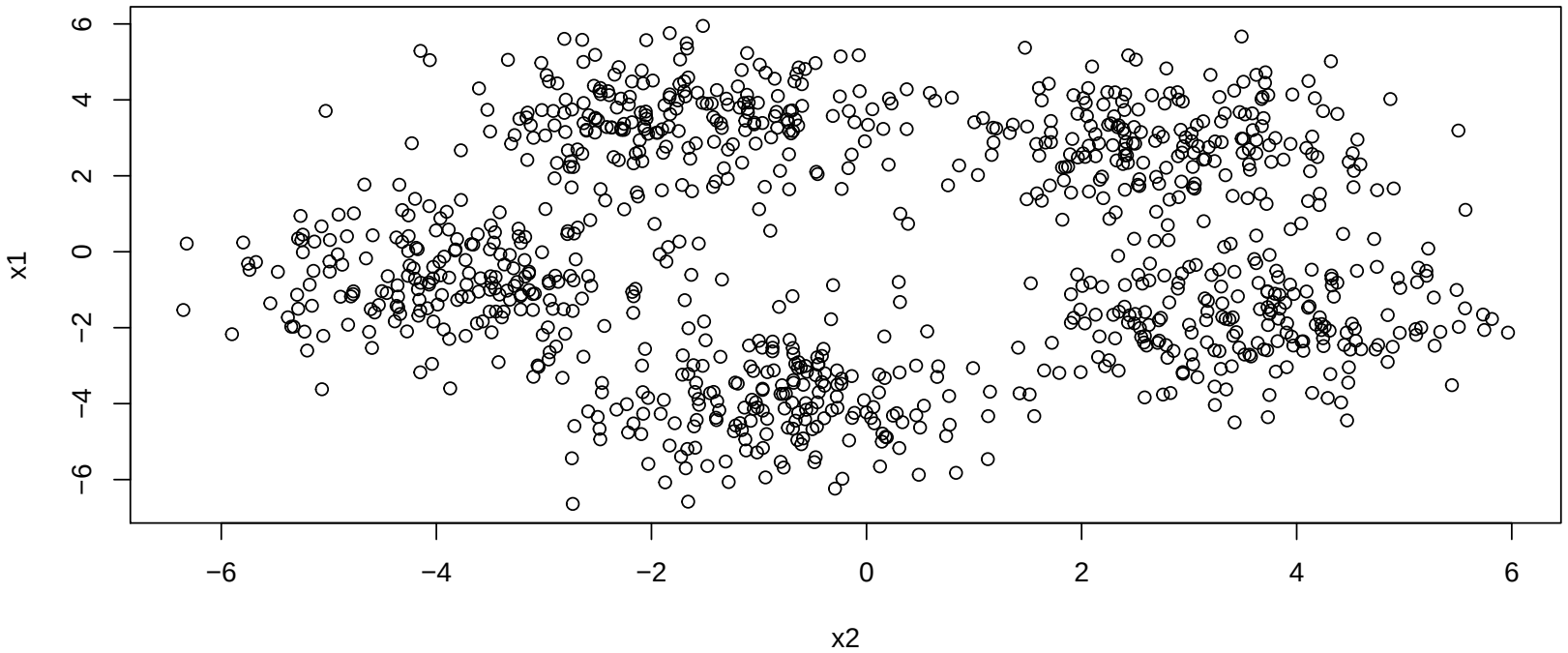
Algorytm *k-średnich* - 2. iteracja 3. krok



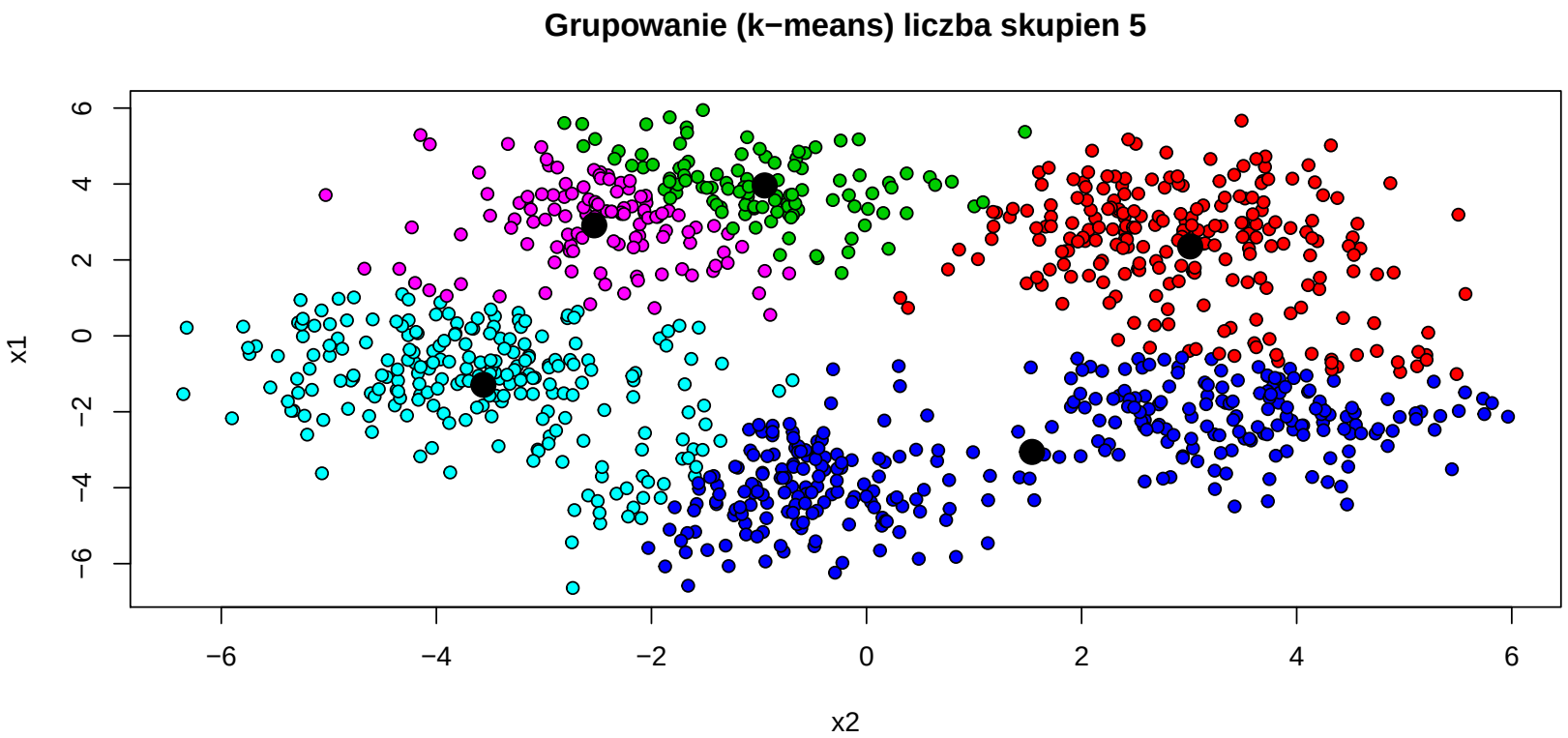
Algorytm *k-średnich* - koniec



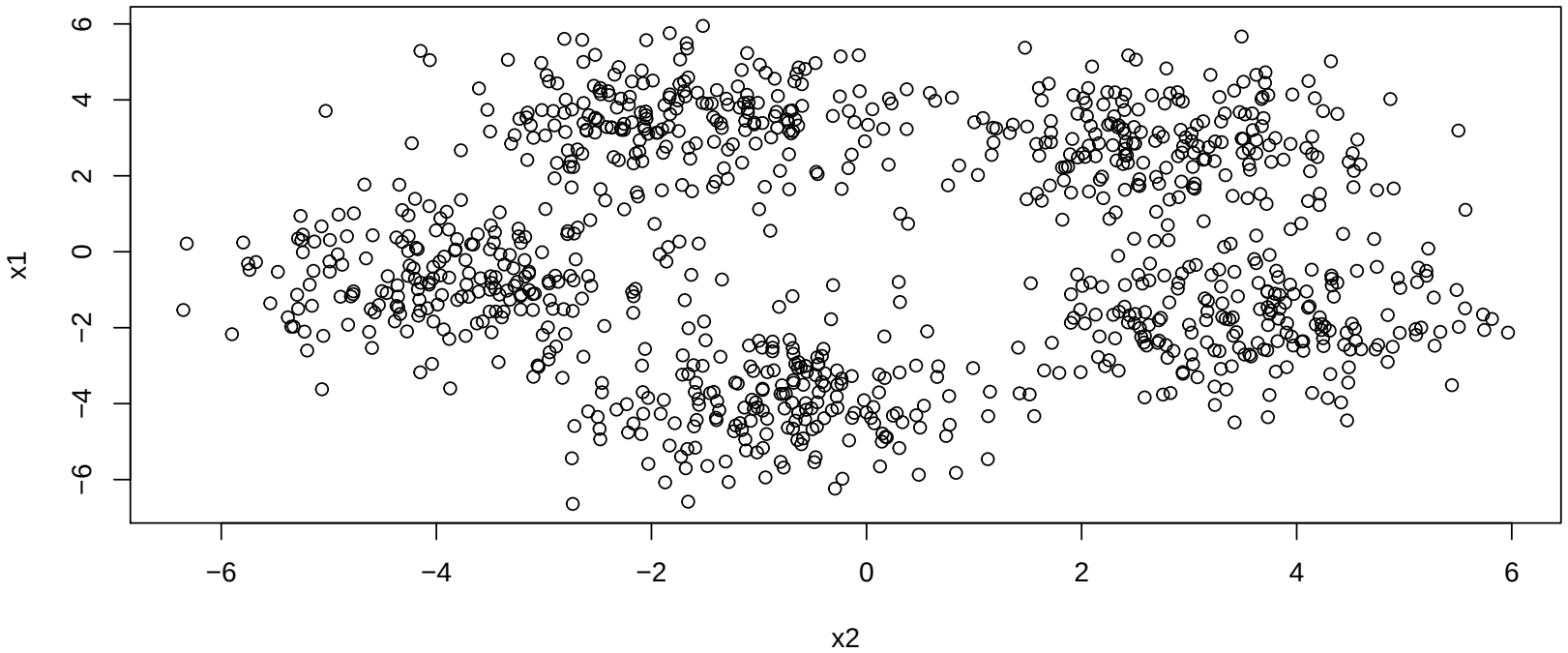
Algorytm *k-średnich* - problem wyboru liczby skupień



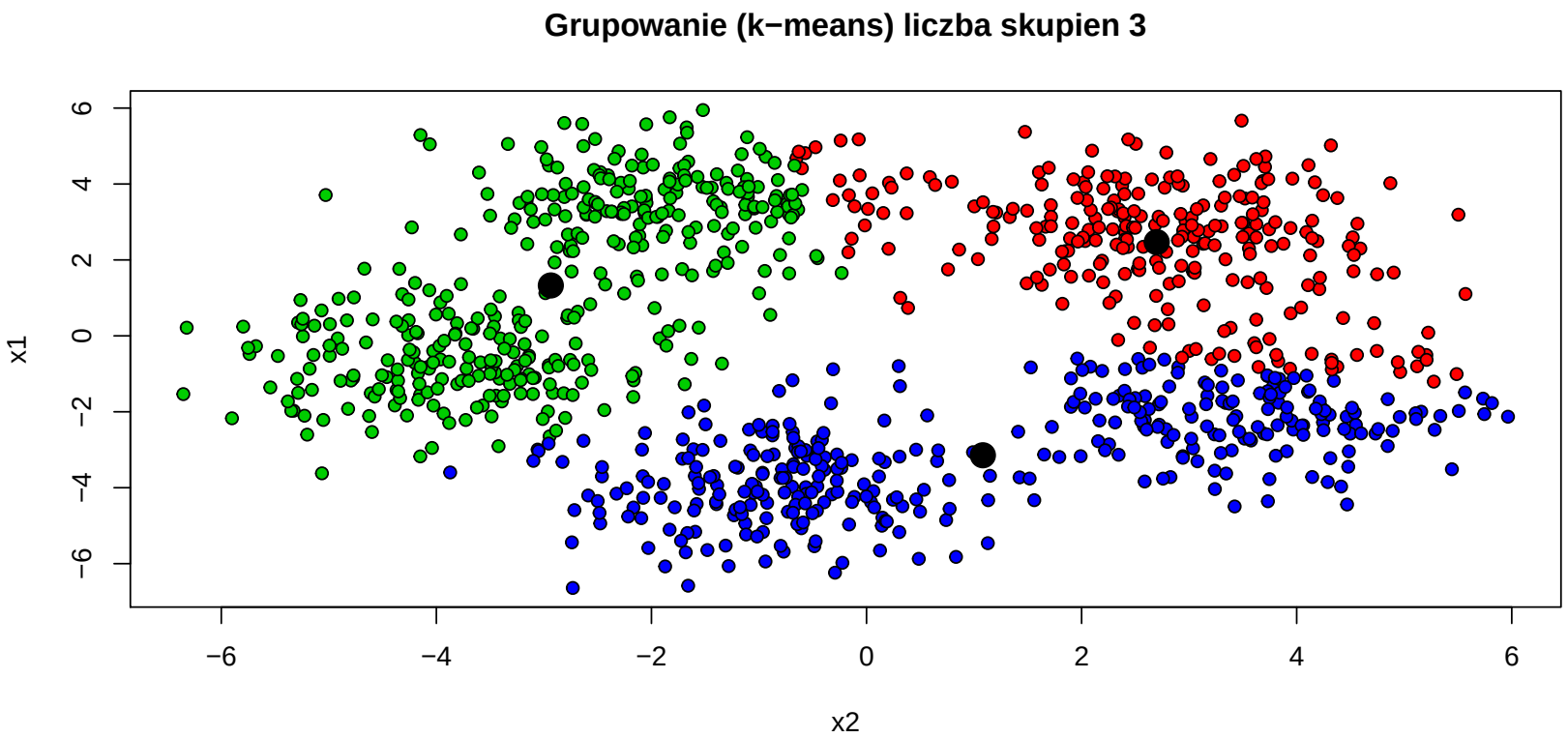
Algorytm *k-średnich* - problem wyboru liczby skupień



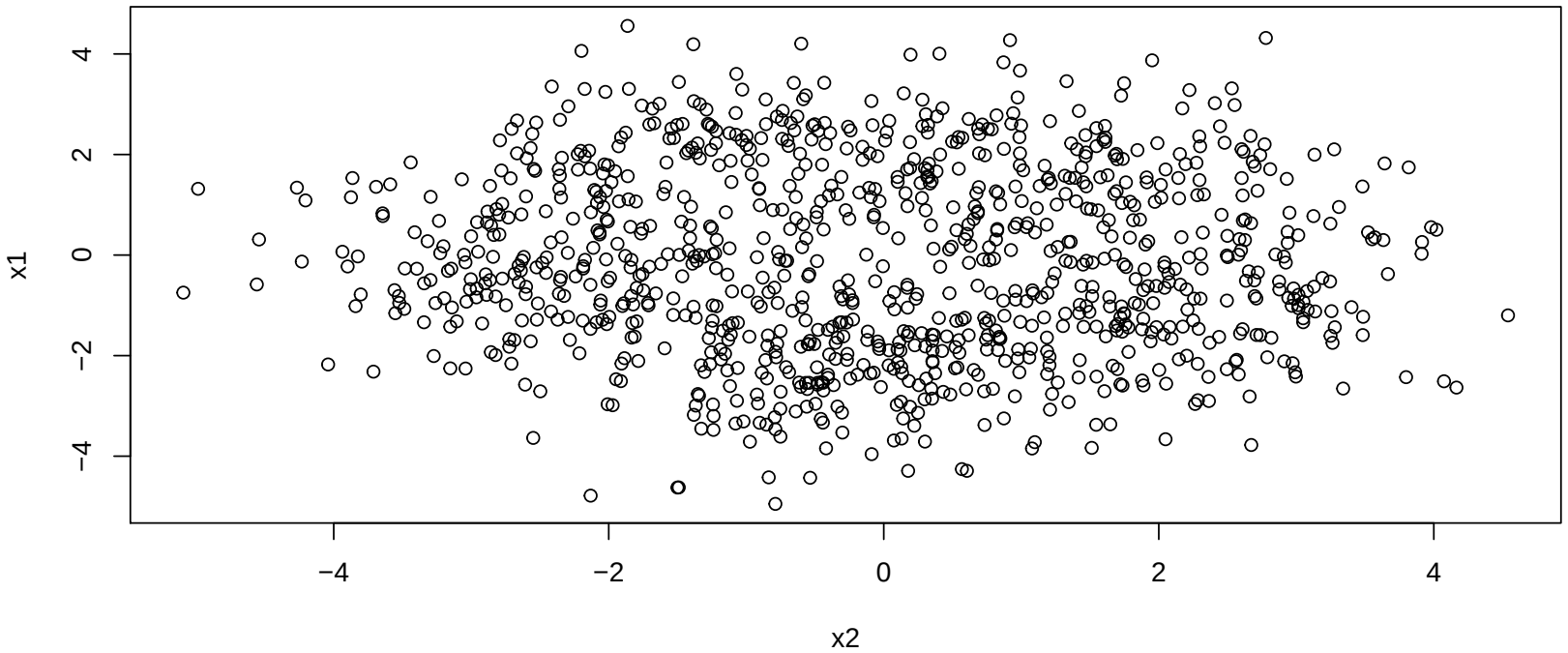
Algorytm *k-średnich* - problem wyboru liczby skupień



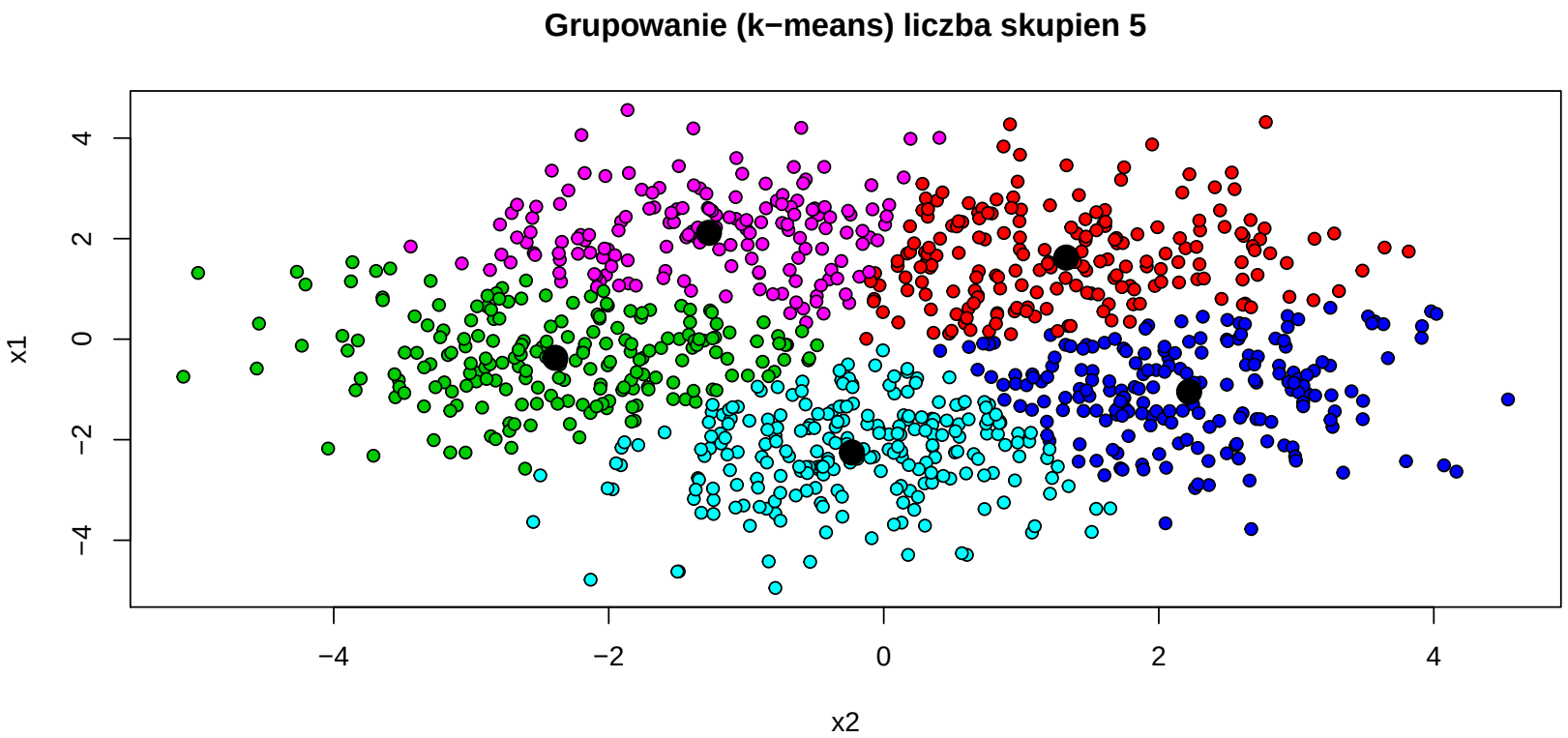
Algorytm *k-średnich* - problem wyboru liczby skupień



Algorytm *k-średnich* - problem wyboru liczby skupień



Algorytm *k-średnich* - problem wyboru liczby skupień



Algorytm *k-średnich* - problem wyboru liczby skupień